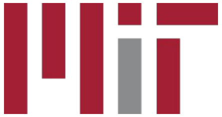
Completeness, Recall and Negation in Open-World Knowledge Bases

Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, Fabian Suchanek

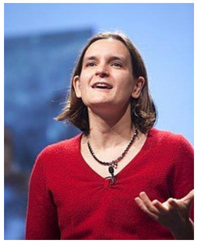
1. Introduction and Foundations (Simon)
2. Predictive Recall Assessment (Fabian)
3. Counts from Text and KB (Shrestha)
4. Identifying Salient Negations (Hiba)
5. Wrap-up (Simon)

Introduction

What common relation ties entities on the right to the entity on the left?



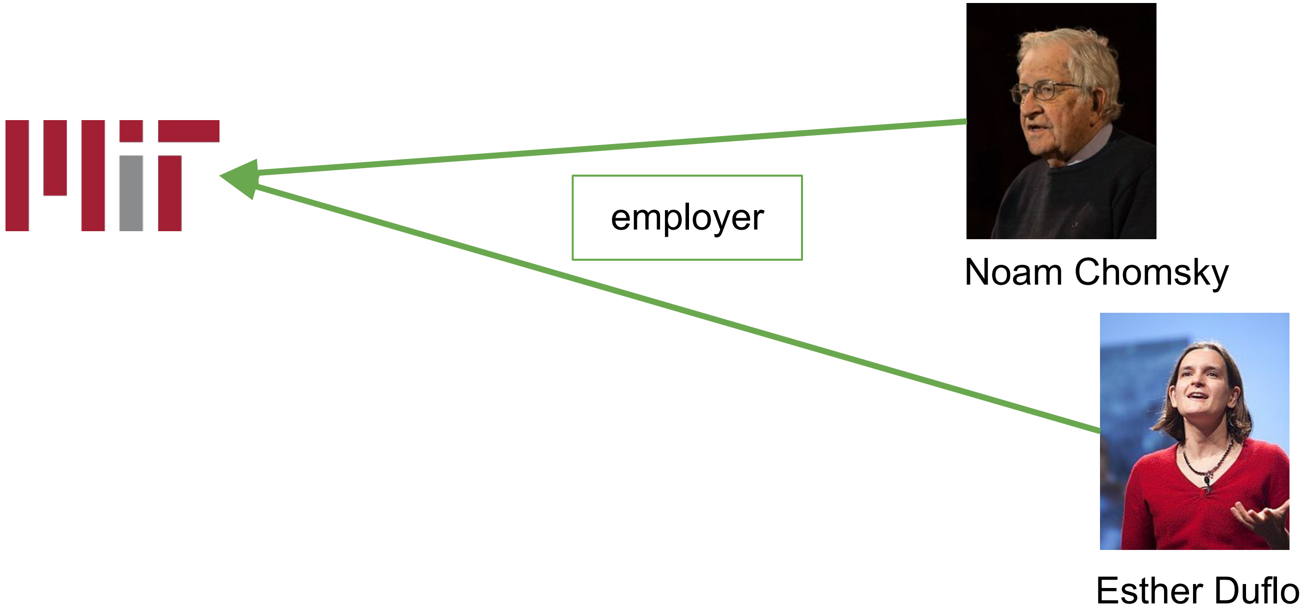
Noam Chomsky



Esther Duflo

Introduction

What common relation ties entities on the right to the entity on the left?



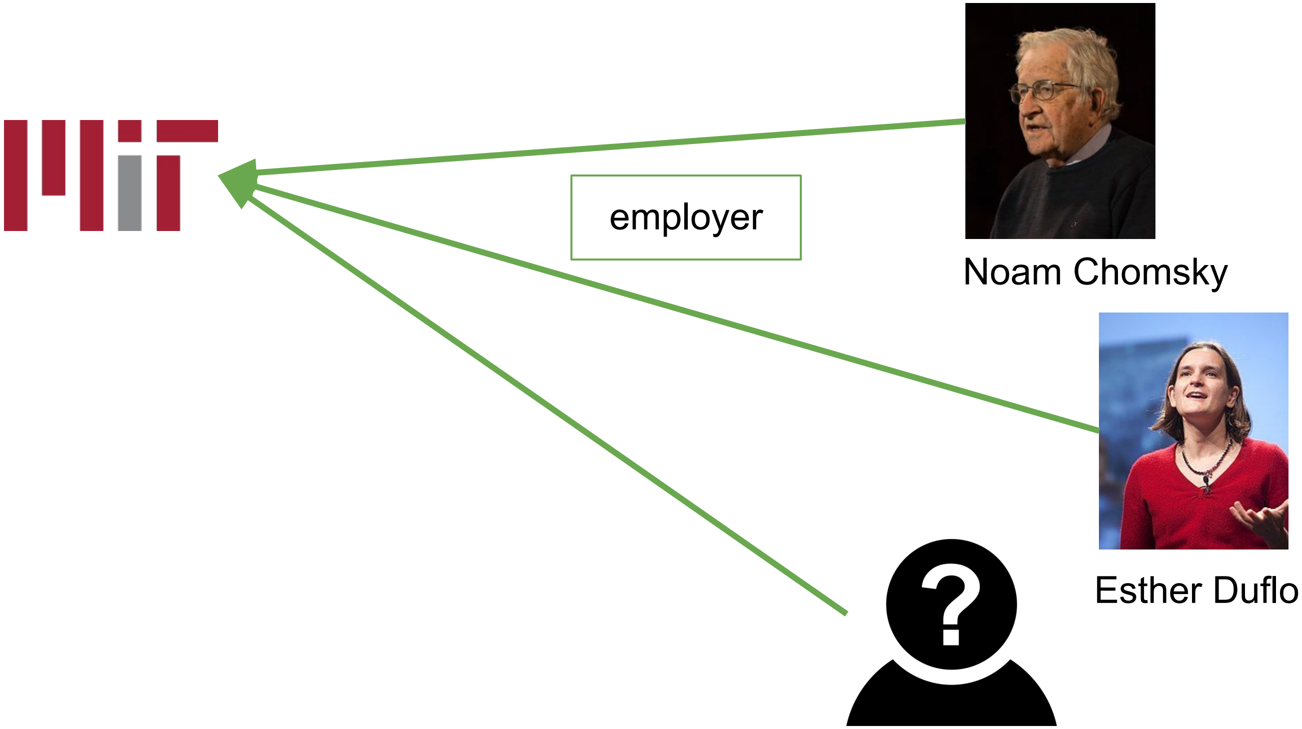
employer

Noam Chomsky

Esther Duflo

Introduction

How many employees does MIT have?



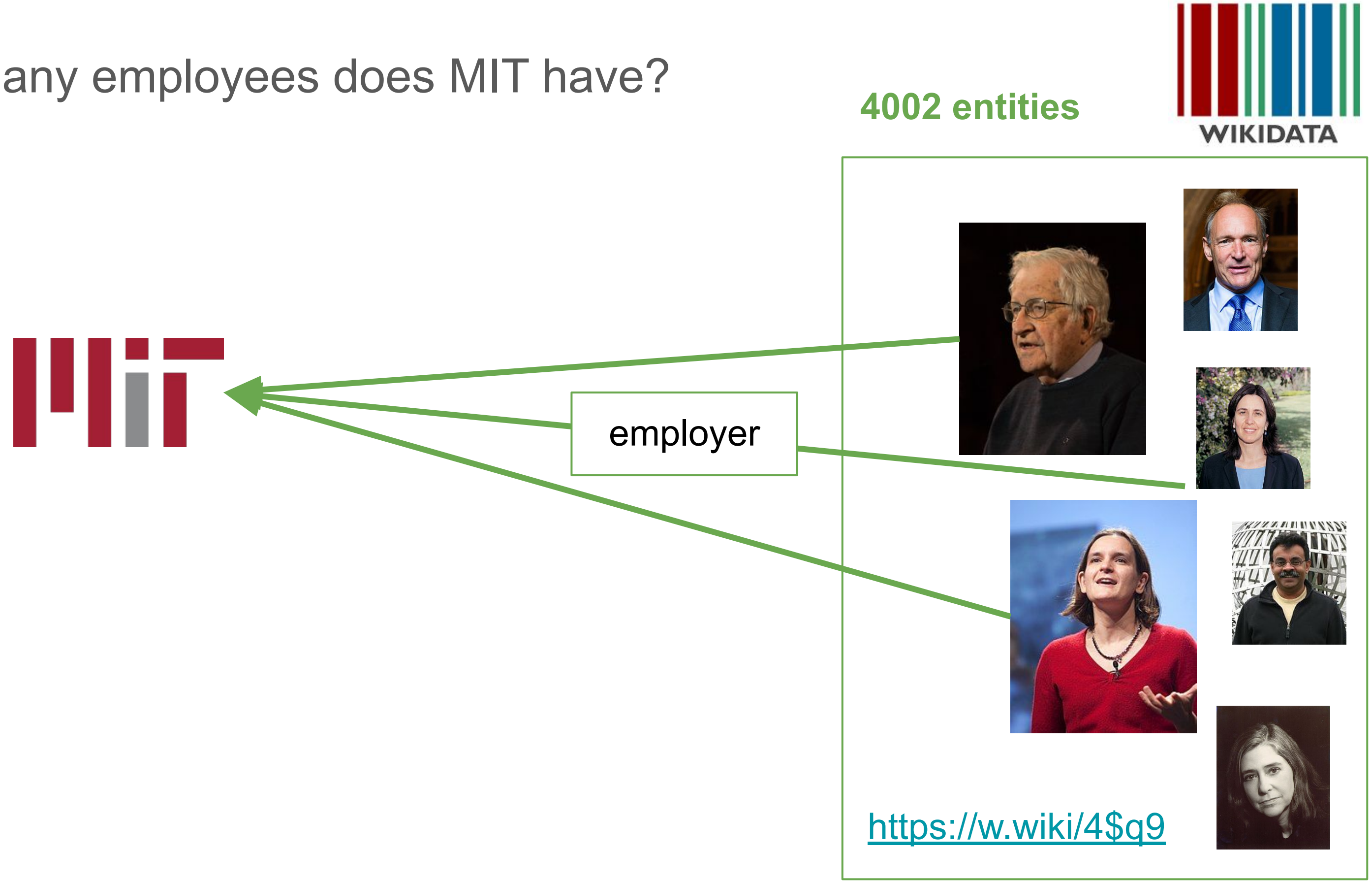
employer

Noam Chomsky

Esther Duflo

Introduction

How many employees does MIT have?



4002 entities

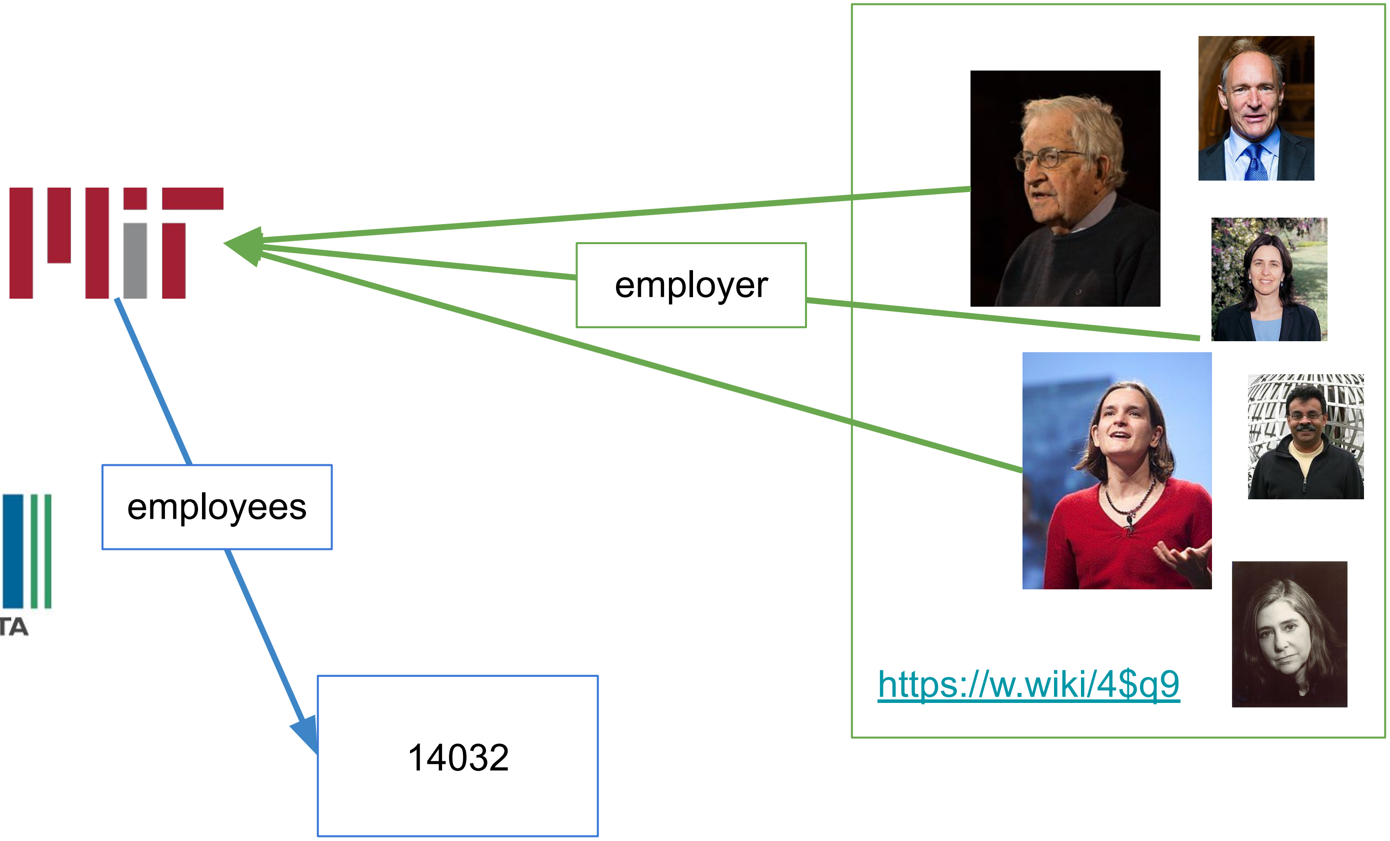
employer

https://w.wiki/4$q9

Introduction

How many employees does MIT have?

4002 entities



employer

employees

14032

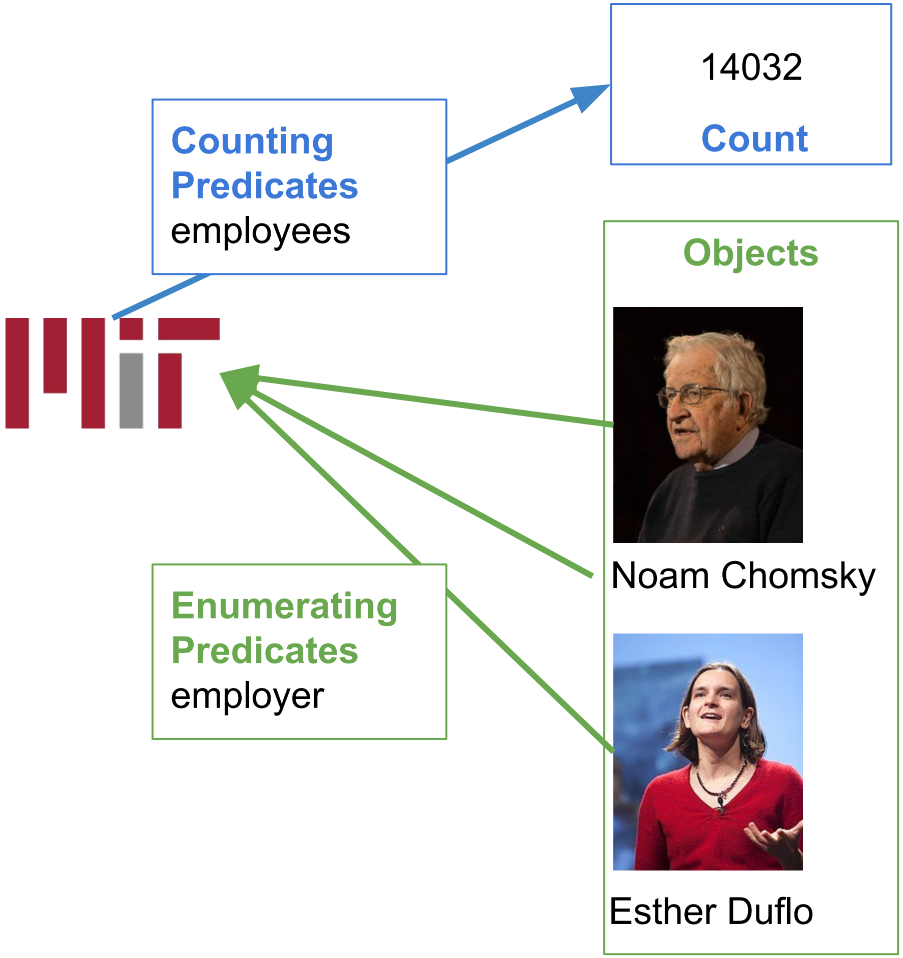
https://w.wiki/4$q9

https://w.wiki/4$qB

Also ..

Introduction

Count Information: Relation between an entity and a set of entities



Counting Predicates
employees

14032
Count

Enumerating Predicates
employer

Objects

Noam Chomsky

Esther Duflo

Expressed as **count** or cardinality of the set

Expressed as **entities** or objects in the set

1. Utility of count information
2. Extracting count information from text
3. Count information in KB
4. How much count information is accounted for?

Utility: Recall assessment

Only entities

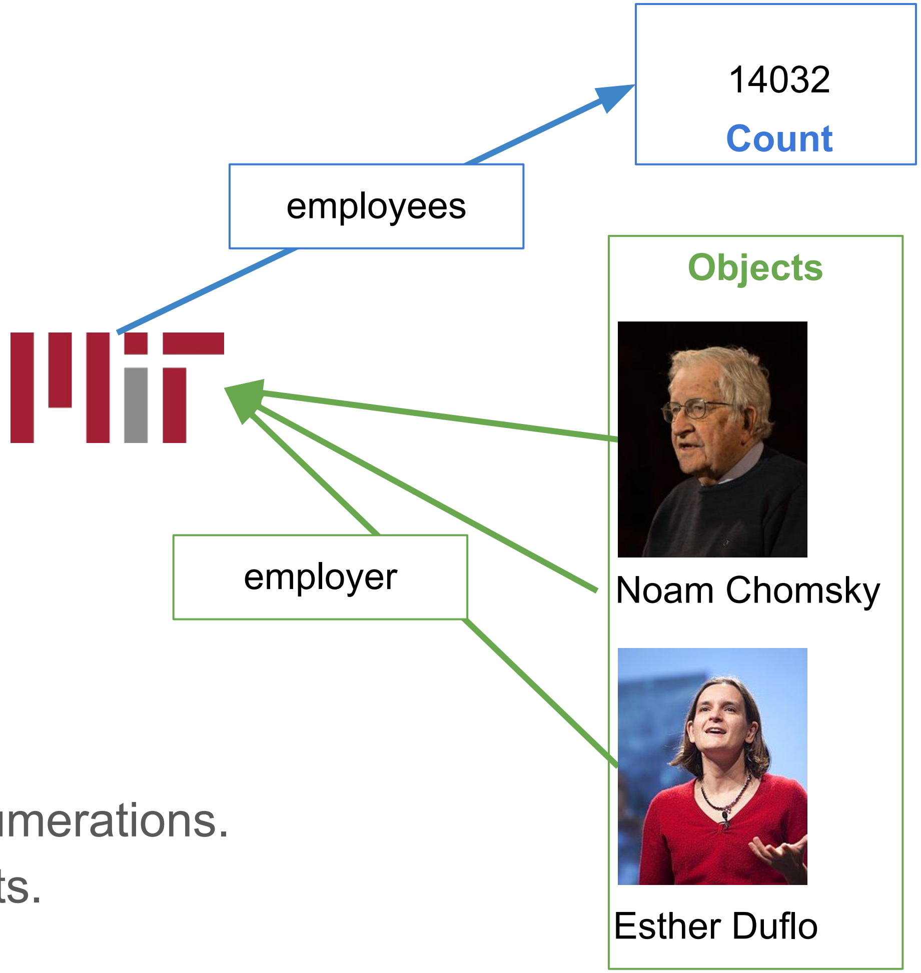
(?x, employer, MIT)

returns a handful of names from KB

Only counts

(MIT, employees, ?y)

gives no insight about the entities

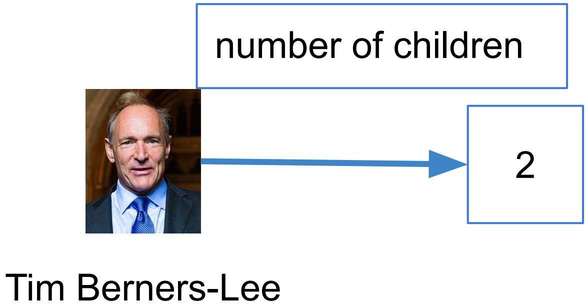


Count and Entities

- Counts enhance incomplete entity enumerations.
- Representative entities enhance counts.

Utility: Recall assessment

KB mixes counts with standard facts



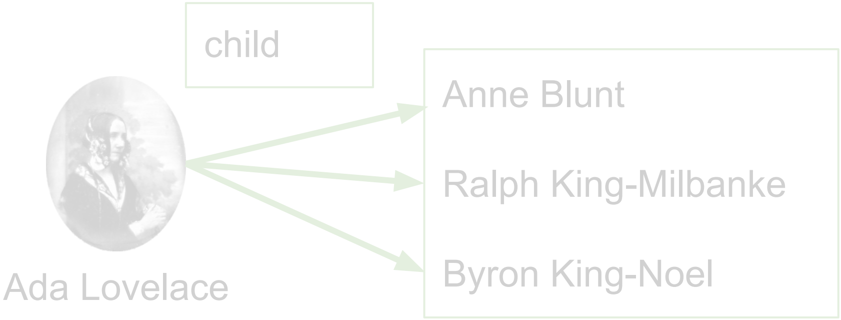
number of children

2

Tim Berners-Lee

How many children does Tim Berners-Lee have?

2 (KB fact)



child

Anne Blunt

Ralph King-Milbanke

Byron King-Noel

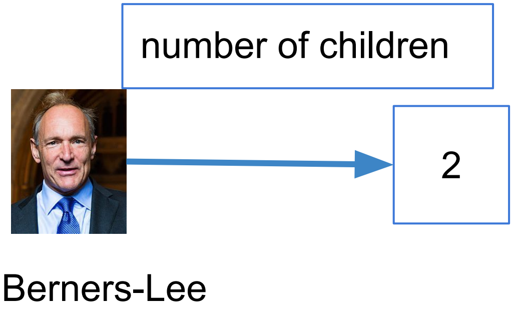
Ada Lovelace

How many children did Ada Lovelace have?

3 (Maybe?)

Utility: Recall assessment

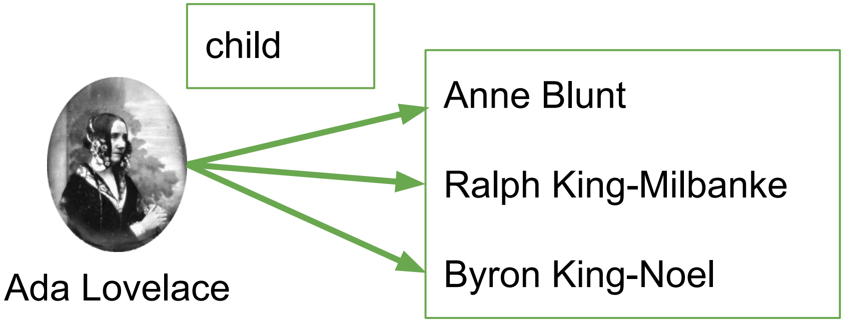
KB mixes counts with standard facts



number of children

2

Tim Berners-Lee



child

Anne Blunt

Ralph King-Milbanke

Byron King-Noel

Ada Lovelace

How many children does Tim Berners-Lee have?

2 (KB fact)

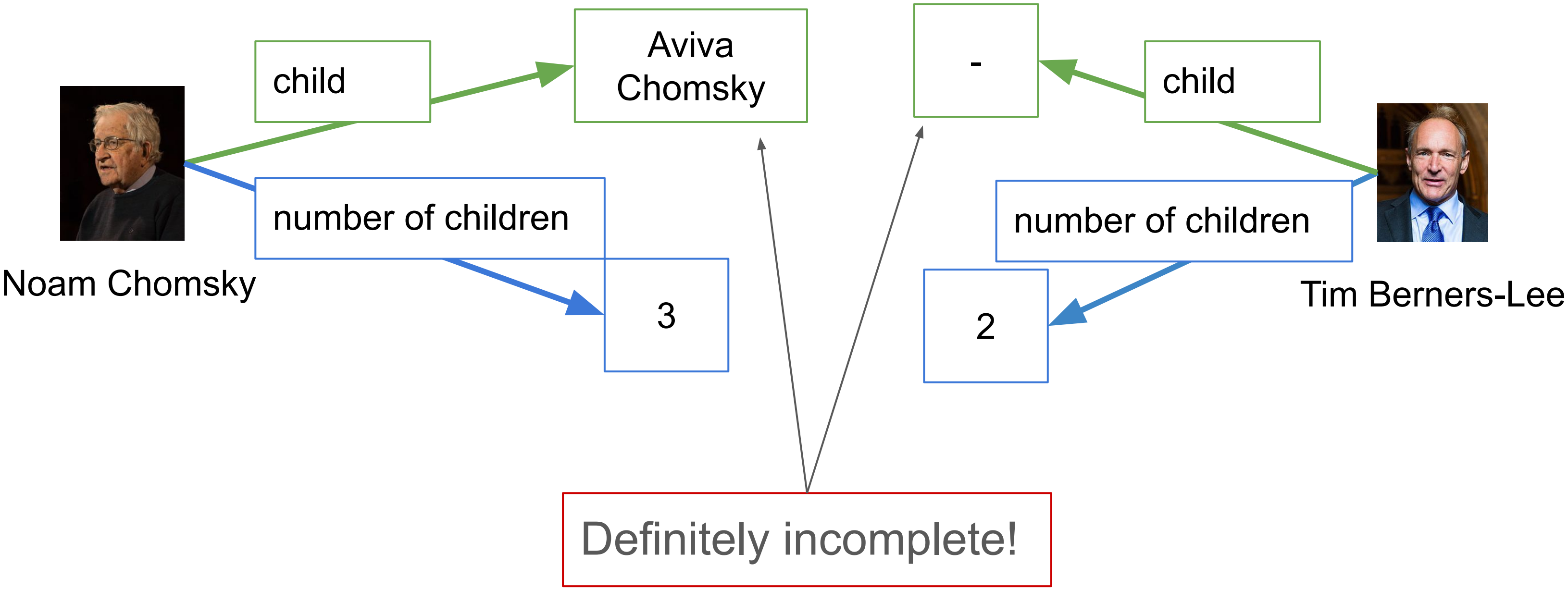
How many children did Ada Lovelace have?

3 (Maybe?)

Enumeration is often of known entities

Utility: Recall assessment

Count information can highlight KB inconsistencies



Noam Chomsky

child

Aviva Chomsky

number of children

3

Tim Berners-Lee

child

-

number of children

2

Definitely incomplete!

1. Utility of count information
2. Extracting count information from text
3. Count information in KB
4. How much count information is accounted for?

Count information from text

Problem: Counting Quantifier Extraction

Input:

- a text about a subject S
- a predicate P

Task: Determine the number of objects in which S stands in relation with P

Subject: Noam Chomsky
Predicate: number_of_children

Chomsky was married to Carol. They had three children together

3

Paramita Mirza, Simon Razniewski, Fariz Darari, Gerhard Weikum
Enriching Knowledge Bases with Quantifiers
International Semantic Web Conference (ISWC) 2018.

Count information from text

Task 1: Identify the count tokens and the compositional cues.

Sequence Labelling of tokens in a sentence on subject S and predicate P with:

- COUNT - for counts
- COMP - for compositional cues
- O - all other tokens

Subject: Noam Chomsky
Predicate: number_of_children

Chomsky	was	married	to	Carol.	They	had	three	children	together
O	O	O	O	O	O	O	COUNT	O	O

Count information from text

Task 1: Identify the count tokens and the compositional cues.

Sequence Labelling of tokens in a sentence on subject S and predicate P with:

- COUNT - for counts
- COMP - for compositional cues
- O - all other tokens

Subject: Angelina Jolie
Predicate: number_of_children

Jolie	has	three	sons	and	three	daughters.
O	O	COUNT	O	COMP	COUNT	O

Count information from text

Task 2: Consolidate count tokens

Return a single answer per text, given subject-predicate pair

1. **Sum up compositional cues**

6

Jolie brought her six children: twins , one daughter and three adopted children to the gala.

Subject: Angelina Jolie
Predicate: number_of_children

Count information from text

Task 2: Consolidate count tokens

Return a single answer per text, given subject-predicate pair

1. Sum up compositional cues
2. **Select prediction per type**

6 (cardinal) **6** (cardinal)

Jolie brought her six children: twins , one daughter and three adopted children to the gala.

Subject: Angelina Jolie
Predicate: number_of_children

6 (cardinal)

Count information from text

Task 2: Consolidate count tokens

Return a single answer per text, given subject-predicate pair

1. Sum up compositional cues
2. Select prediction per type
3. **Rank mention types**

cardinal >> number-related terms >> ordinals >> indefinite article

two children >> twins >> second child >> a child

Count information from text

Task 2: Consolidate count tokens

Return a single answer per text, given subject-predicate pair

1. Sum up compositional cues
2. Select prediction per type
3. **Rank mention types**

cardinal >> number-related terms >> ordinals >> indefinite article
two children >> twins >> second child >> a child

Jolie brought her six children: twins , one daughter and three adopted children to the gala.

Subject: Angelina Jolie
Predicate: number_of_children

6 (cardinal)

Count information from text

Ground Truth

Use KB information as Ground Truth

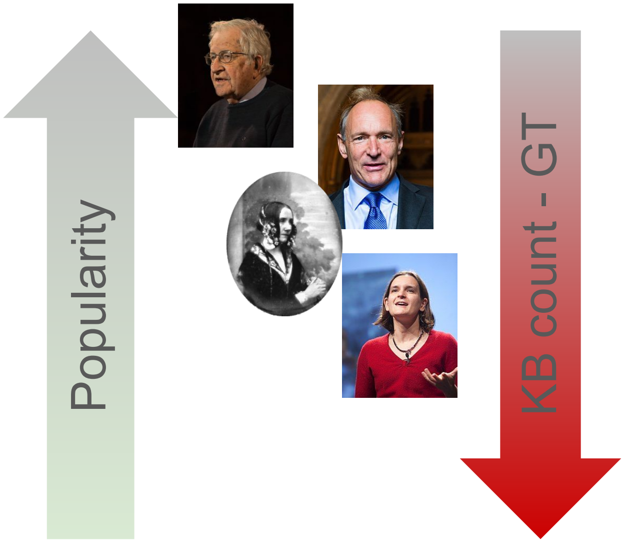
Challenges

KB incompleteness negatively impacts training quality

Solution

Consider only popular KB entities

Set upper bound for predicate count value = 99th percentile of KB predicate value distribution



1. Utility of count information
2. Extracting count information from text
3. Count information in KB
4. How much count information is accounted for?

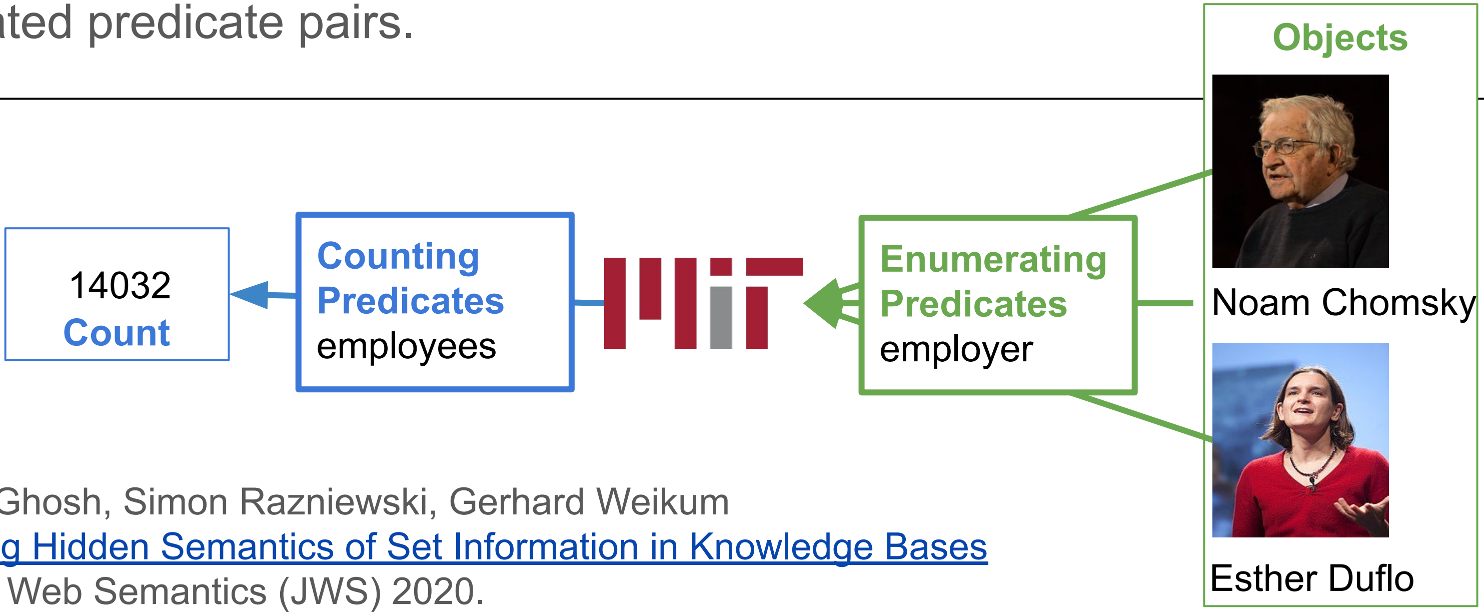
Count information in KB

Problem: Identification of semantically related count predicates

Input:

- a set of KB triples *(s,p,o)*
- and its inverse predicate triples *(s,p^{-1},o)*

Task: Determine counting and enumerating predicates and semantically related predicate pairs.



Objects

14032
Count

Counting Predicates
employees

Enumerating Predicates
employer

Noam Chomsky

Esther Duflo

Shrestha Ghosh, Simon Razniewski, Gerhard Weikum
Uncovering Hidden Semantics of Set Information in Knowledge Bases
Journal of Web Semantics (JWS) 2020.

Count information in KB

Task 1: Identification of the count predicates - counting and enumerating

KB predicates

academic_staff, staff, faculty

number_of_children

wins, doubles_titles, singles_titles

...

work_institution^{-1}, workplace^{-1}, work_institutions^{-1}

child

gold^{-1}

Count information in KB

Task 1: Identification of the two variants of count predicates

Counting Predicates

academic_staff, staff, faculty

number_of_children

...

wins, doubles_titles, singles_titles

Enumerating Predicates

work_institution^{-1}, workplace^{-1}, work_institutions^{-1}

child

...

gold^{-1}

Challenge:

- The separation is not clear.
- Not all counting predicates store (single) integers
- Not all enumerating predicates store entities

Count information in KB

Task 1: **Identification** of the two variants of count predicates

Counting Predicates

academic_staff, staff, faculty

number_of_children

...

wins, doubles_titles, singles_titles

Enumerating Predicates

work_institution^{-1}, workplace^{-1}, work_institutions^{-1}

child

...

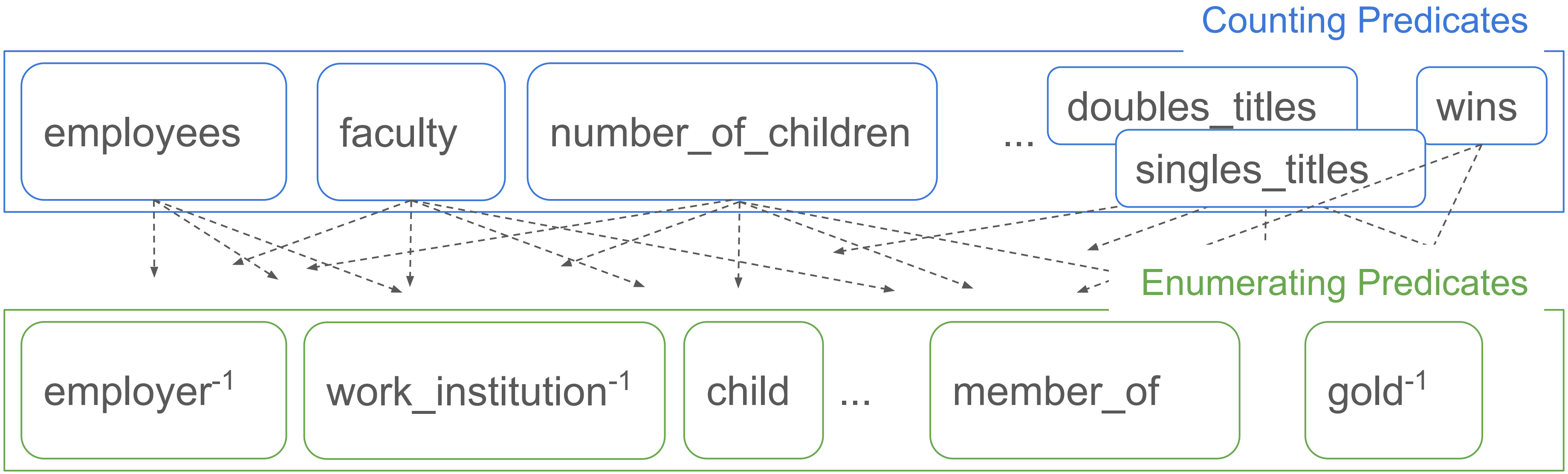
gold^{-1}

Supervised Classification using:

- **Textual Features** - count predicates are more often used in plural form
- **Type Information** - classes of subject and objects
- **KB statistics** - #objects per subject, datatype distribution of the objects

Count information in KB

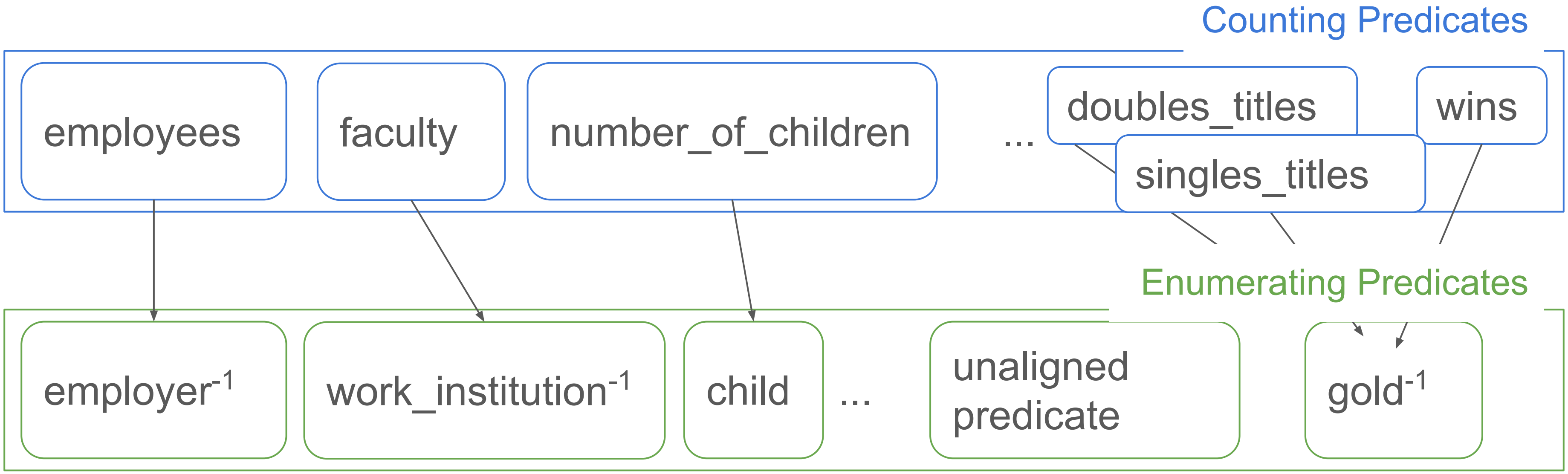
Task 2: **Aligning** pairs of counting and enumerating predicates



Challenge: KB facts are sparse and unclean.

Institutions can use faculty_size, employees or staff to mean the same thing.

Count information in KB



Heuristics used for the predicate pair ***(e,c)***, where ***e*** stores entities and ***c*** counts.

1. Predicate pair co-occurrences - #subjects ***e*** and ***c*** co-occur
2. Value distribution - number of objects of ***e*** compared to count in ***c***
   a. is it equal for all subjects?
   b. is there any correlation?
3. Linguistic similarity - do ***e*** and ***c*** talk share topical similarity?

1. Utility of count information
2. Extracting count information from text
3. Count information in KB
4. How much count information is accounted for?

How much count information is accounted for?

Counts from text

173k new count facts increasing KB knowledge by **77%**

from just 4 Wikidata properties across 10 classes

2,205 negative assertions

2.5M new count facts increasing KB knowledge by **28.3%**

How much count information is accounted for?

Counts from text

173k new count facts increasing KB knowledge by **77%**

from just 4 Wikidata properties across 10 classes

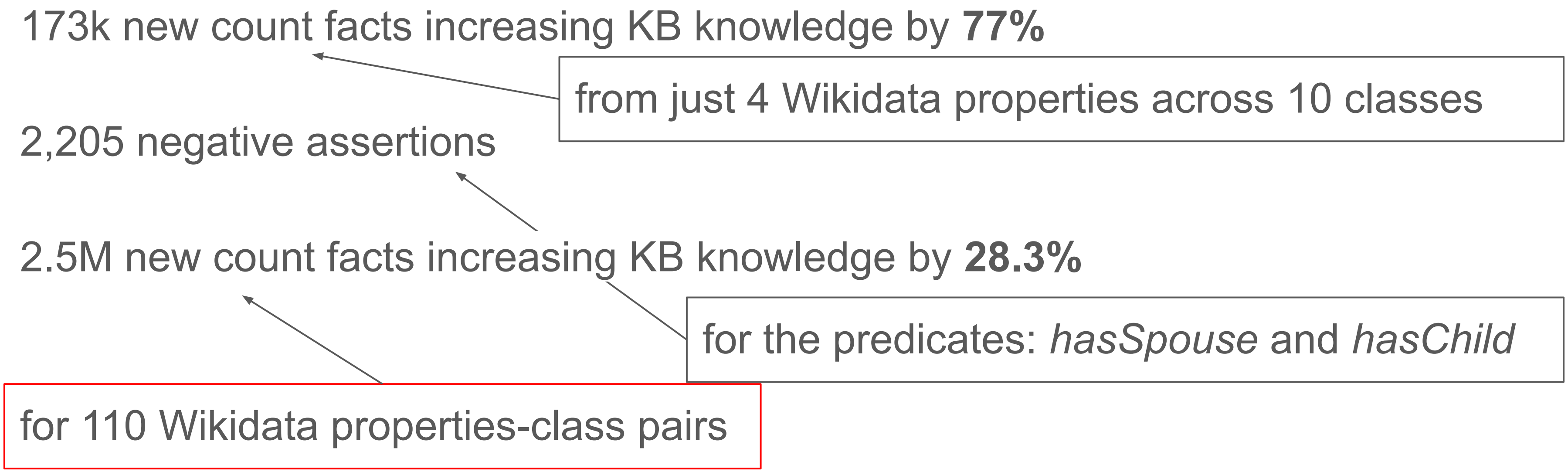
2,205 negative assertions

2.5M new count facts increasing KB knowledge by **28.3%**

for the predicates: *hasSpouse* and *hasChild*

How much count information is accounted for?

Counts from text



173k new count facts increasing KB knowledge by **77%**

from just 4 Wikidata properties across 10 classes

2,205 negative assertions

2.5M new count facts increasing KB knowledge by **28.3%**

for the predicates: *hasSpouse* and *hasChild*

for 110 Wikidata properties-class pairs

Paramita Mirza, Simon Razniewski, Fariz Darari, Gerhard Weikum
Enriching Knowledge Bases with Quantifiers
International Semantic Web Conference (ISWC) 2018.

How much count information is accounted for?

KB	Enumerating
DBpedia-raw	4,090
DBpedia mapped	308
Wikidata-truthy	203
Freebase	7,614
Total	12,215

Number of predicted enumerating KB predicates

From more than 36k frequent predicates across KBs including inverses.

How much count information is accounted for?

KB	Enumerating	Counting
DBpedia-raw	4,090	5,853
DBpedia mapped	308	898
Wikidata-truthy	203	1,067
Freebase	7,614	1,687
Total	12,215	9,505

Number of predicted counting KB predicates

From more than 26k frequent predicates across KBs.

How much count information is accounted for?

Number of predicted count predicates and KB alignments

KB	**Enumerating**	**Counting**	**Alignments**
DBpedia-raw	4,090	5,853	3,703
DBpedia mapped	308	898	270
Wikidata-truthy	203	1,067	31
Freebase	7,614	1,687	274
Total	12,215	9,505	4,278

Quite a low number of alignments: indicative of KB sparsity

Shrestha Ghosh, Simon Razniewski, Gerhard Weikum
Uncovering Hidden Semantics of Set Information in Knowledge Bases
Journal of Web Semantics (JWS) 2020.

Summary

Count information in the KB (Ghosh et al. JWS 2020)

- Exists as integers (counting) and set of entities (enumerating)
- Are semantically related
- Can be used for recall assessment, QA and KB curation

Count information in text (Mirza et al. ACL 2017)

- Is linguistically diverse
- Can be used for populating KBs.

Other works have explored

- Embedding cardinality constraints in link predictors (Munoz et al. SIGAPP 2018).
- Enhancing KB-QA with count informations (Ghosh et al. ESWC 2020).
- Numerical commonsense knowledge in LMs (Lin et al. EMNLP 2020).
- Answering count queries across multiple text sources (Ghosh et al. SIGIR 2022).

Summary

Count information in the KB (Ghosh et al. JWS 2020)

- Exists as integers (counting) and set of entities (enumerating)
- Are semantically related
- Can be used for recall assessment, QA and KB curation

Count information in text (Mirza et al. ACL 2017)

- Is linguistically diverse
- Can be used for populating KBs.

Other works have explored

- Embedding cardinality constraints in link predictors (Munoz et al. SIGAPP 2018).
- Enhancing KB-QA with count informations (Ghosh et al. ESWC 2020).
- Numerical commonsense knowledge in LMs (Lin et al. EMNLP 2020).
- Answering count queries across multiple text sources (Ghosh et al. SIGIR 2022).

Summary

Count information in the KB (Ghosh et al. JWS 2020)

- Exists as integers (counting) and set of entities (enumerating)
- Are semantically related
- Can be used for recall assessment, QA and KB curation

Count information in text (Mirza et al. ACL 2017)

- Is linguistically diverse
- Can be used for populating KBs.

Other works have explored

- Embedding cardinality constraints in link predictors (Munoz et al. SIGAPP 2018).
- Enhancing KB-QA with count informations (Ghosh et al. ESWC 2020).
- Numerical commonsense knowledge in LMs (Lin et al. EMNLP 2020).
- Answering count queries across multiple text sources (Ghosh et al. SIGIR 2022).

Takeaways: Counts from text and KB

1. Count information
   - Is a relation between an entity and a set of entities
   - Expressed in counts and entities
   - Occurs as semantically related counting and enumerating predicates
   - Is present in KBs and text
2. Utility of count information
   - Recall assessment
   - Enhanced question answering
3. Challenges
   - KBs are inconsistent: mix counts with standard facts
   - KBs are sparse and incomplete
   - Counts in text is linguistically diverse

References

1. Paramita Mirza, Simon Razniewski, Fariz Darari, Gerhard Weikum. Enriching Knowledge Bases with Quantifiers. International Semantic Web Conference (ISWC) 2018.
2. Emir Muñoz, Pasquale Minervini, and Matthias Nickles. Embedding cardinality constraints in neural link predictors. Symposium on Applied Computing (ACM/SIGAPP) 2019.
3. Shrestha Ghosh, Simon Razniewski, Gerhard Weikum. Uncovering Hidden Semantics of Set Information in Knowledge Bases. Journal of Web Semantics (JWS) 2020.
4. Shrestha Ghosh, Simon Razniewski, and Gerhard Weikum. CounQER: A System for Discovering and Linking Count Information in Knowledge Bases. European Semantic Web Conference (ESWC) 2020.
5. Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. Empirical Methods in Natural Language Processing (EMNLP) 2020.
6. Shrestha Ghosh, Simon Razniewski, Gerhard Weikum. Answering Count Queries with Explanatory Evidence. Special Interest Group in Information Retrieval (ACM SIGIR) 2022.